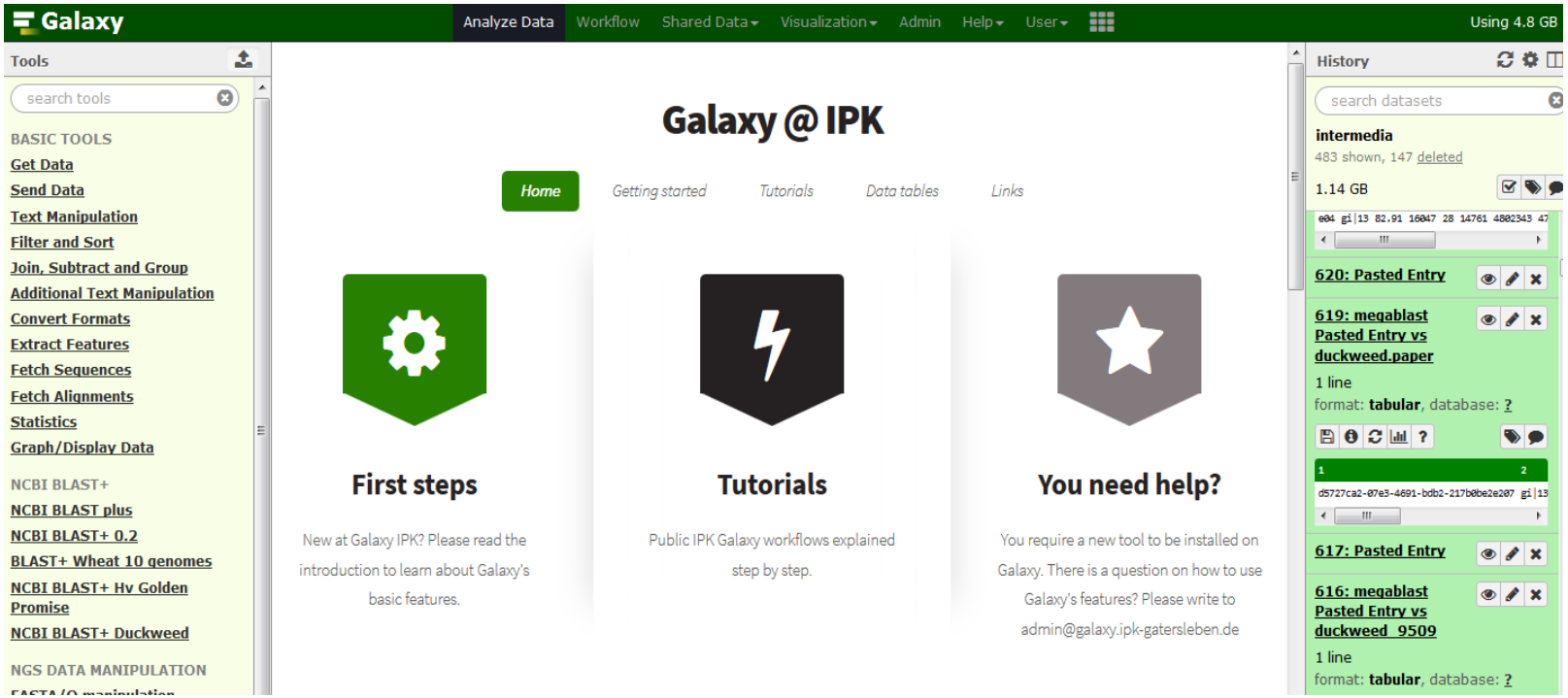
Anne Fiebig, Leibniz-Institut für Pflanzengenetik und Kulturpflanzenforschung (IPK)

Galaxy

Workshop spring school 2019



A short introduction to Galaxy

Agenda

- Introduction of Galaxy‘s main functionality
  - Creating a login for usegalaxy.eu
    https://usegalaxy.eu/join-training/ipk_springschool
- Coffee break
- Hands-on: First steps on usegalaxy.eu
  - Uploading data
  - Running tools
  - Creating a simple workflow
  - Setting up a dataset collection
  - (Run the RNA-Seq workflow)

The Galaxy platform

What is Galaxy?

"Galaxy is an open, web-based platform for accessible, reproducible and transparent computational research."

http://galaxyproject.org

The Galaxy platform

- Developed at Penn State, Johns Hopkins, OHSU and Cleveland Clinic with substantial outside contributions

- **Open source** under Academic Free License
  - 116 public servers (both general-purpose and domain-specific)
  - Numerous internal servers at institutes and companies
  - Citizen science projects

The Galaxy platform

Galaxy-E initiative (https://github.com/65MO/Galaxy-E)

- Initially devoted to citizen science projects related to Biodiversity started in 2015
- When uploading a dataset, its datatype can be project coordinated of the french National Museum of Natural History MNHN called "65 Millions d'observateurs"
  - Bird observation
- Users are directly enrolled in the analysis

Members of the public are invited to take part in a national bird counting survey in their gardens, on January 27 and 28.

For the sixth year running, the biodiversity network the LPO and national museum the *Muséum national d'Histoire naturelle (MNHN)* are joining together in the initiative, which seeks to follow bird populations living in proximity with humans, in a bid to understand their condition and to help develop measures to protect the animals.

Everyone can take part, whether from a city or the countryside, a big garden or small yard, or even from a balcony or windowsill. Even a public park counts.

Only a few words of French (to understand the website when submitting your details) and a connection to the internet are needed to add your observations - there is no need to be a bird or nature expert.

Oldiefan / CC0 Creative Commons // Gerrit Vyn / The Cornell Lab of Orthnithology // Dfaulder / CC BY 2.0

Why using galaxy?

…as a Bioinformatician?

Open Source, python-based downloadable package that can be deployed in individual labs:

- modularized
- add new tools (maintain various versions)
- integrate new data sources
- easely plug in your own components

…as a scientist without bioinformatic background?

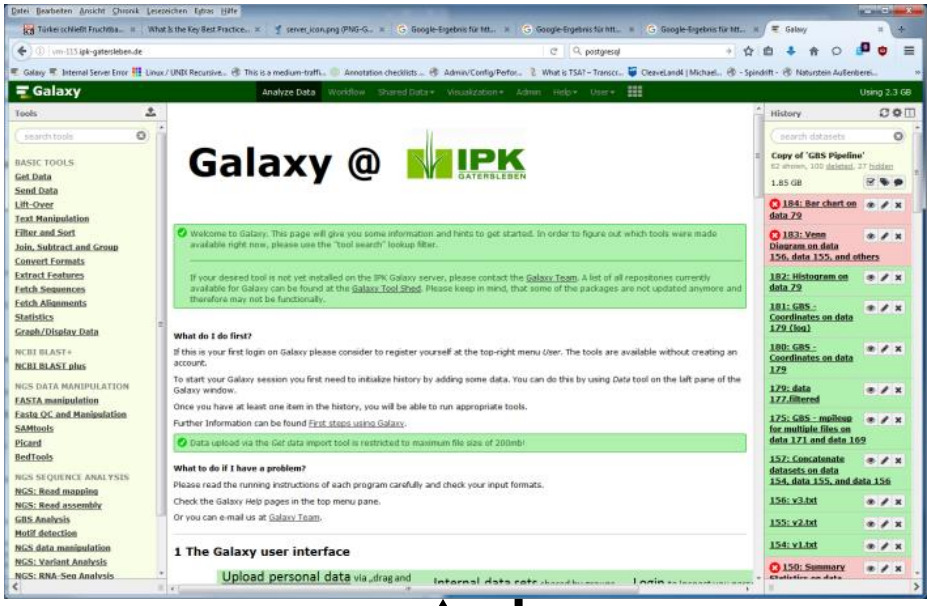
You can:

- analyze genome-scale NGS data without bash scripting
- work with big datasets, genomic regions, sequences etc.
- create and use Galaxy workflows
- share results and workflows with a user or make it available to anyone

Galaxy Workflow

Issues command

View results



Inspect the results

Parsing HTTP request

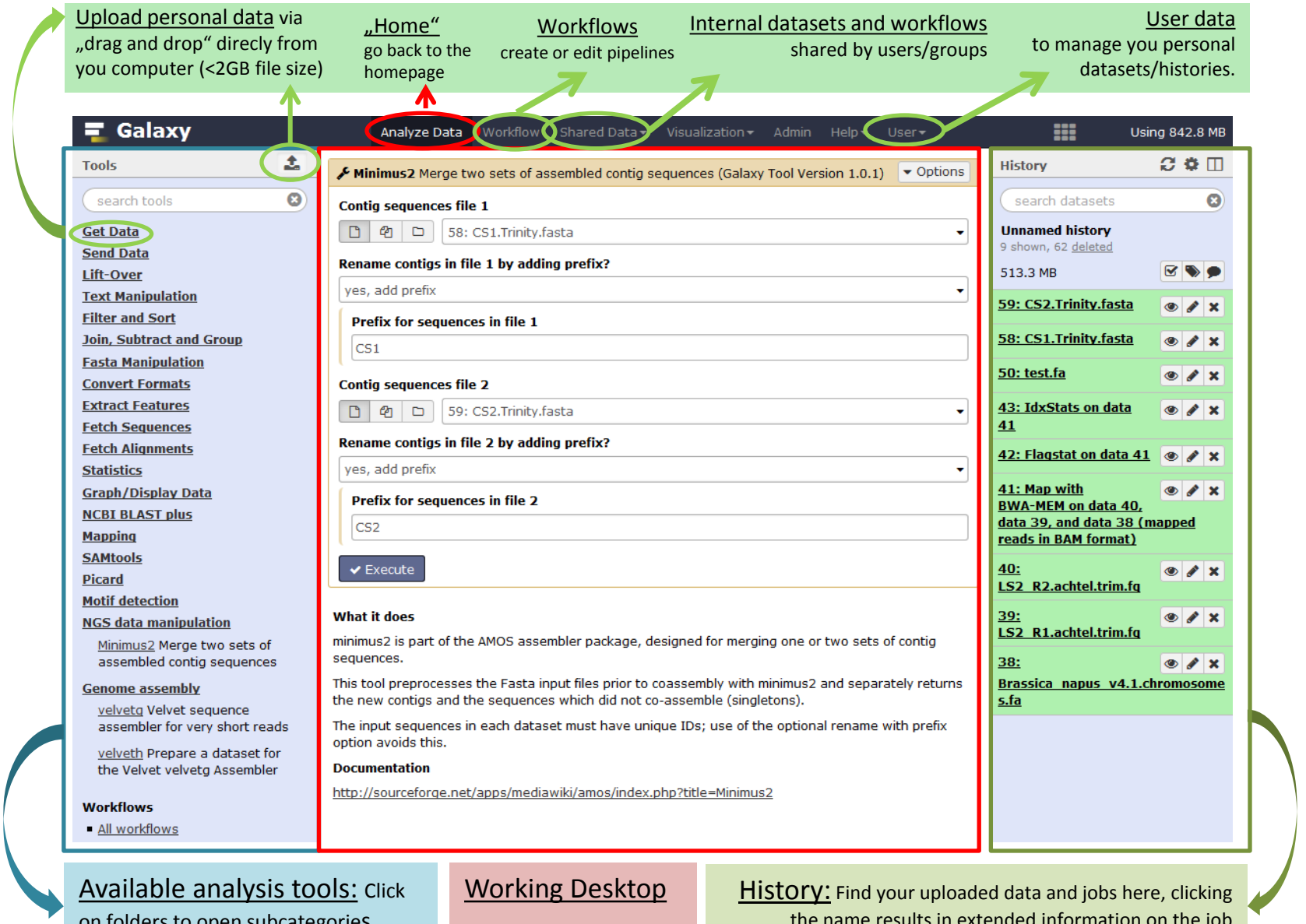
Identifies the tools via ID using the database

Getting tool description (XML)

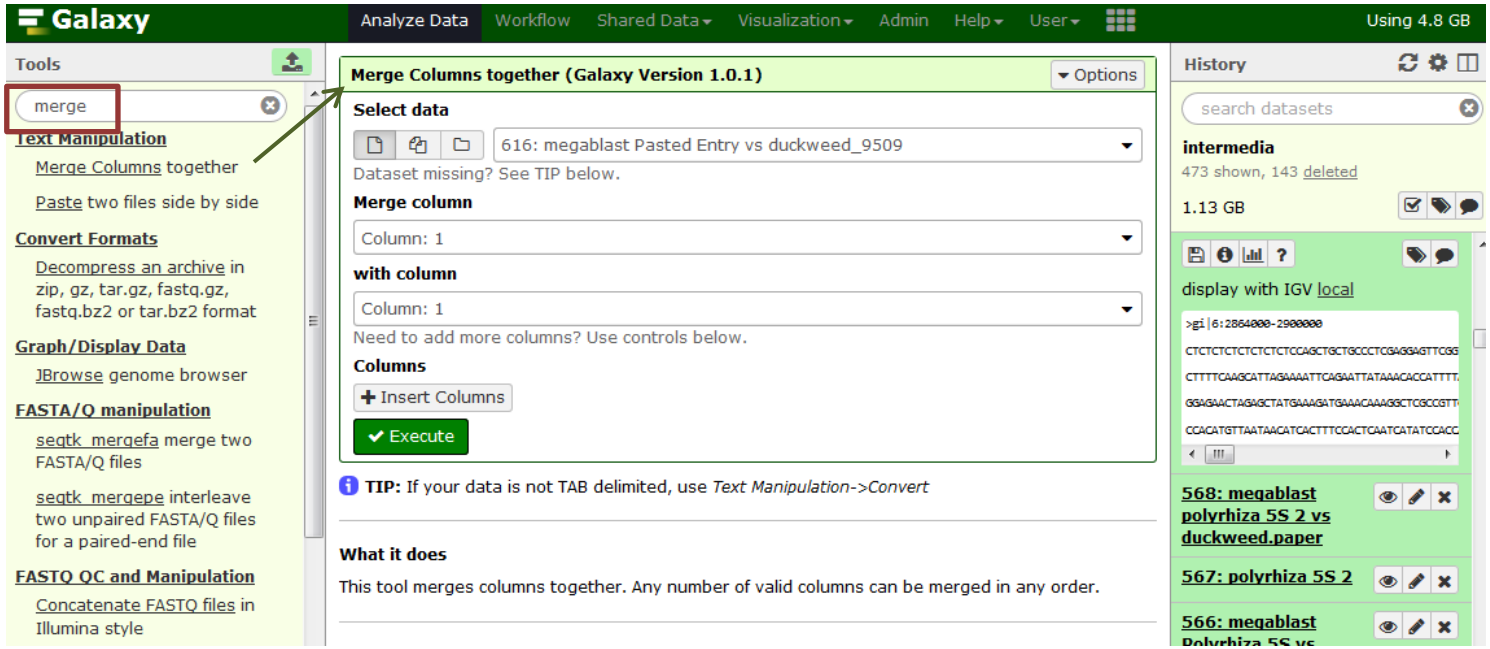
Queues the tool (.sh script in woking directory)

1 Overview

The IPK Galaxy user interface

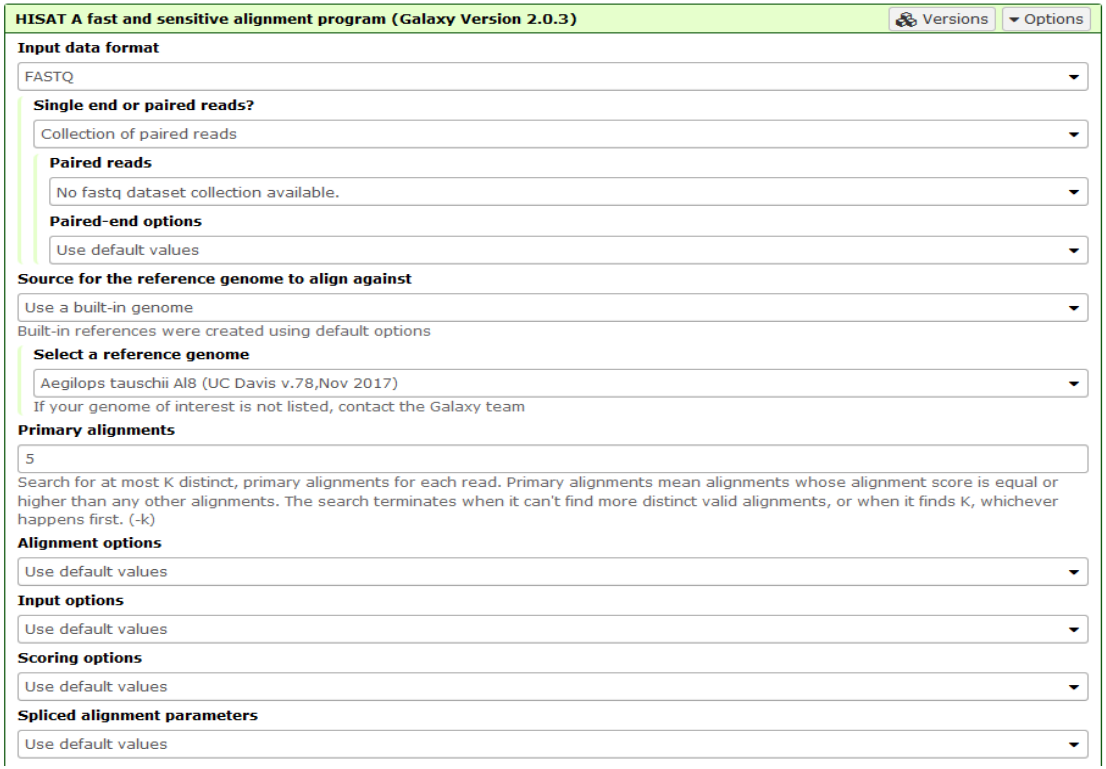


Tools



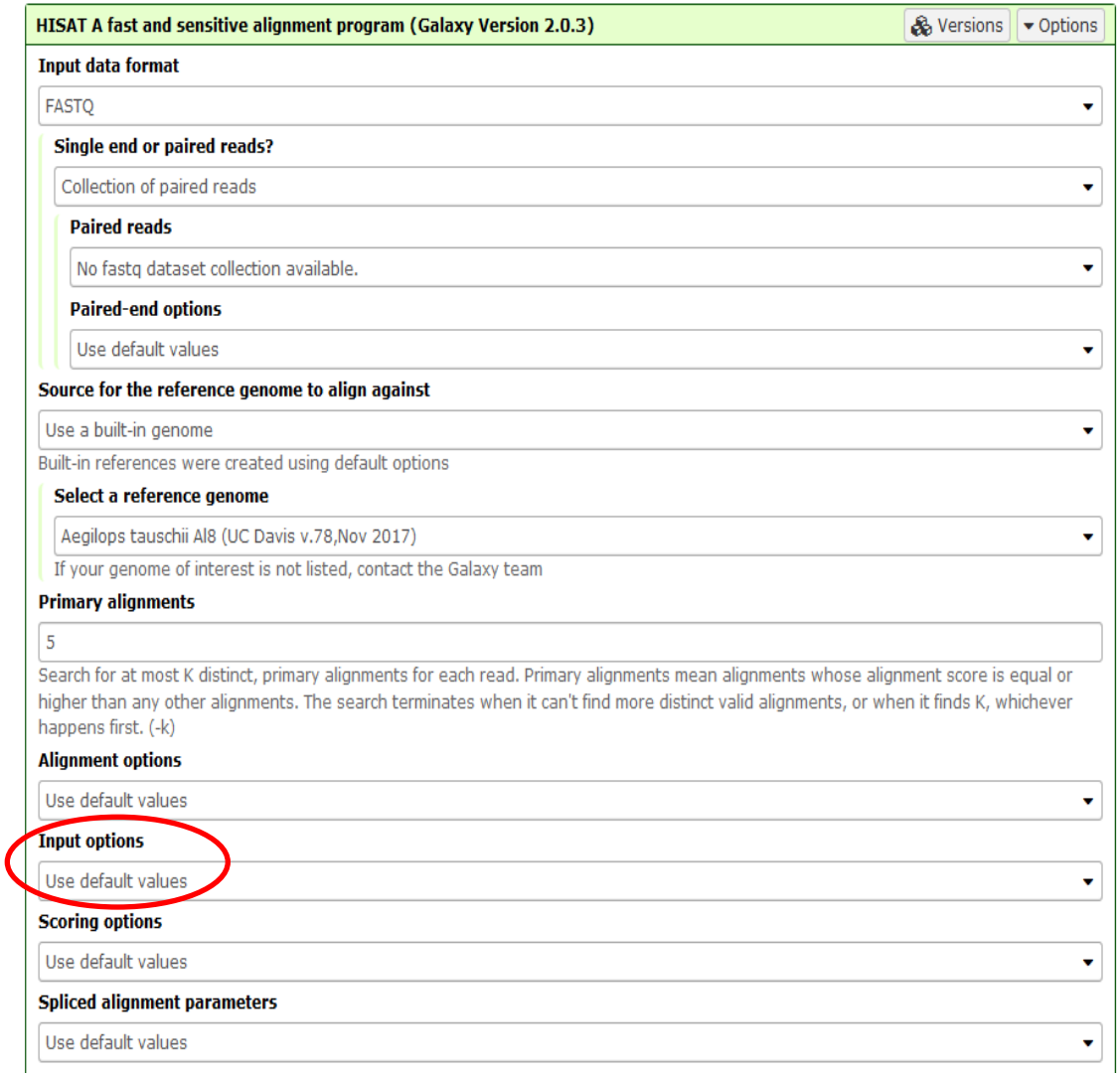
- Organised in categories
- The tool search helps in finding a tool in a crowded toolbox

Tool interface

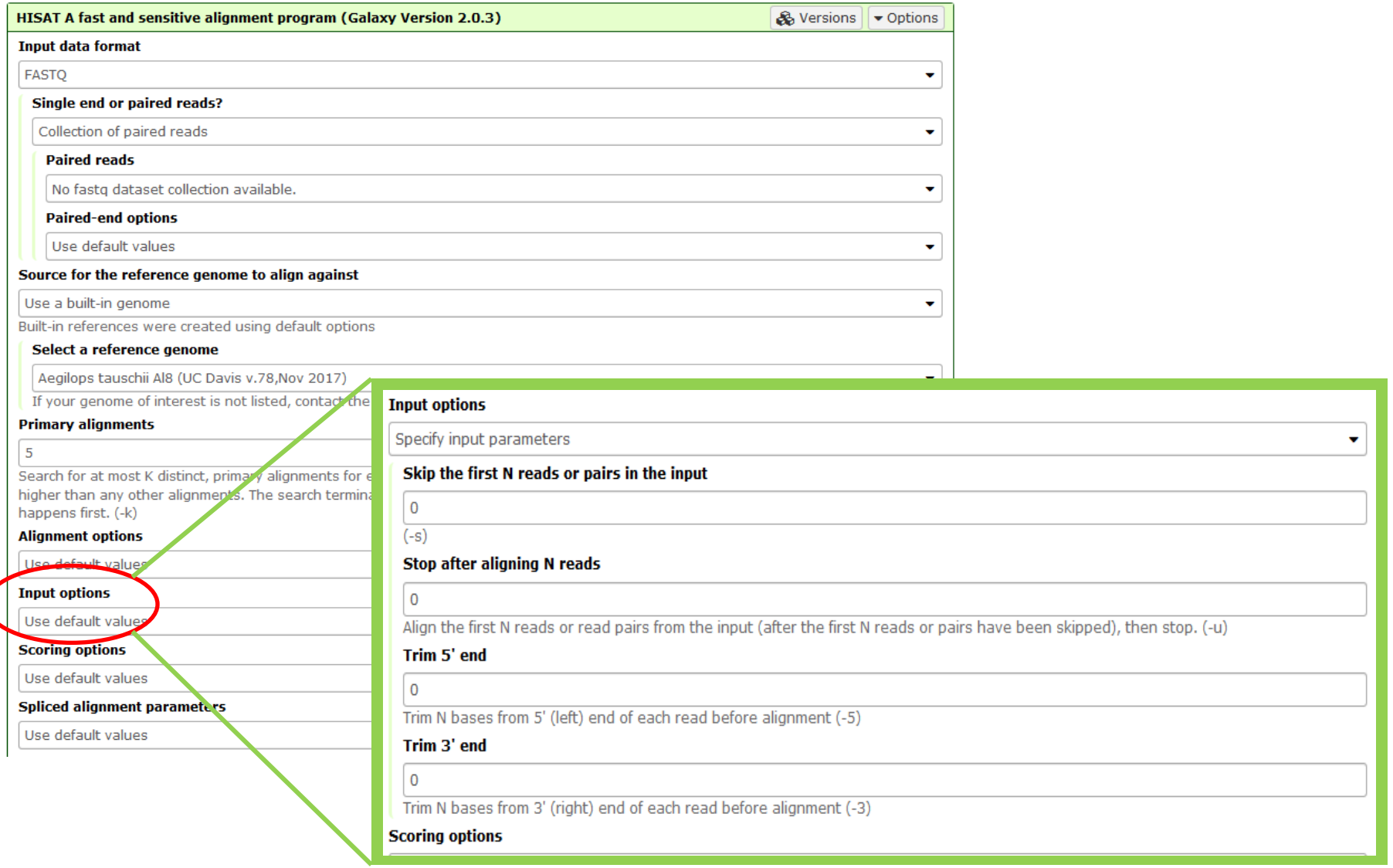


- Tools are simply text files with:
  - input datasets, parameters, commands, and outputs
  - help, tests, citations, dependency requirements
- New versions can be installed without removing old ones to ensure reproducibility

Tool interface



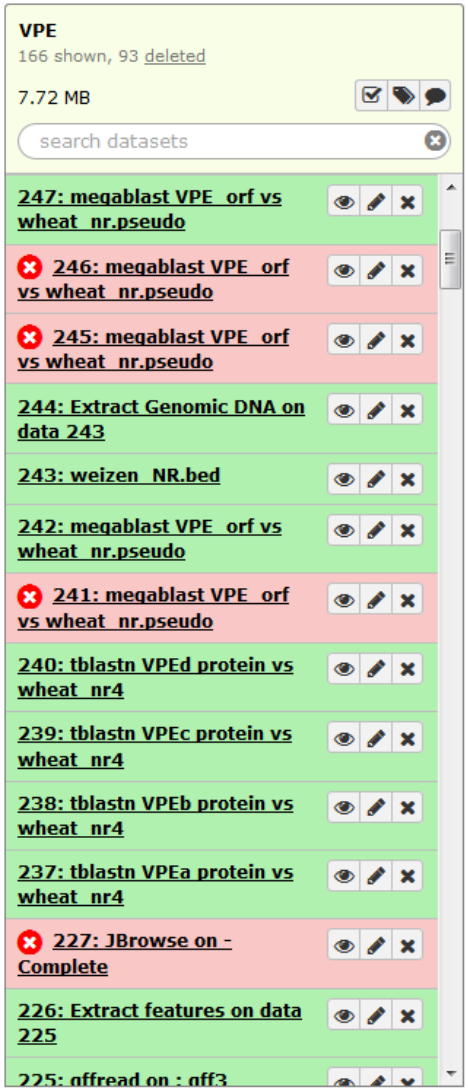
Tool interface



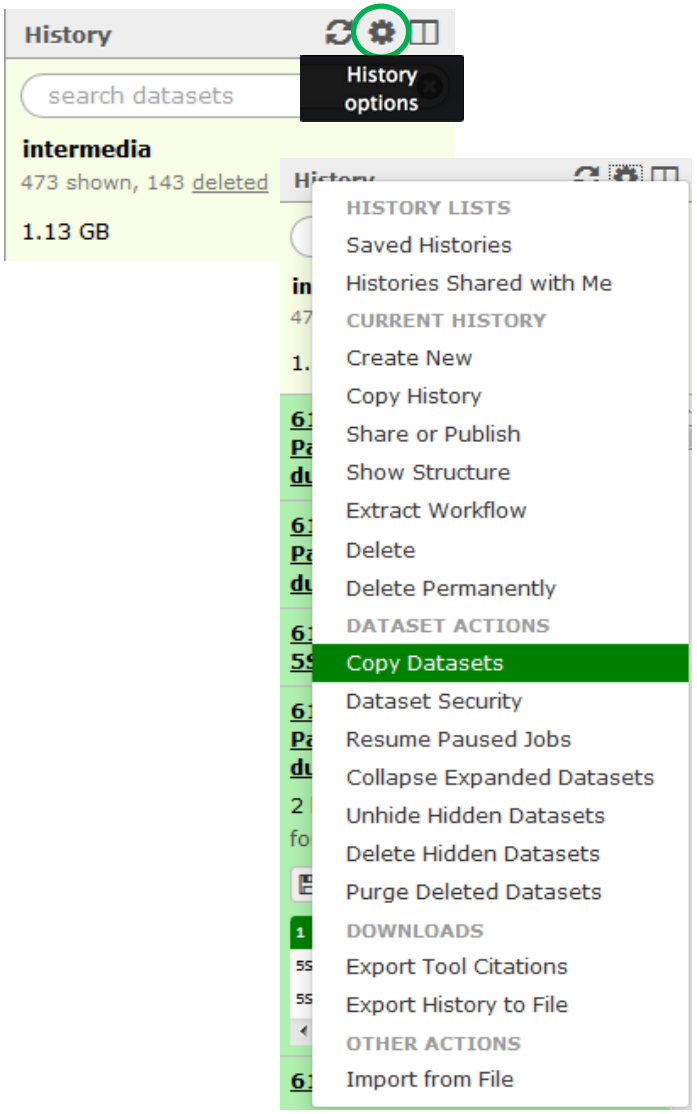
History

- Location of all analyses
  - Collects all datasets produced by tools
  - Collects all operations performed on the data

- For each dataset (the heart of Galaxy's reproducibility), the history tracks
  - Name, format, size, creation time, datatype-specific metadata
  - Tool ID, version, inputs, parameters
  - Standard output (stdout) and error (stderr)
  - State (waiting, running, success, failed)
  - Hidden, deleted, purged



History options

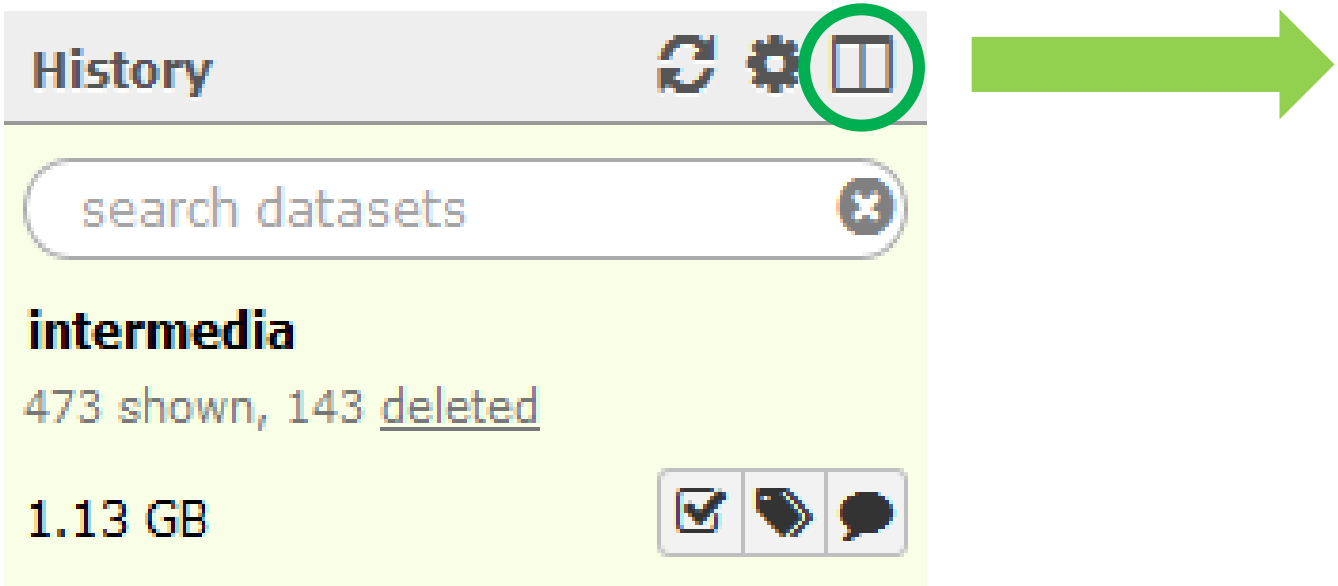


History behavior is controlled by the *History options* (gear icon)

- *Create New* history will **not** make your current history disappear
- *Copy Datasets* from one history to another and save disk space for your quota
- Several delete options

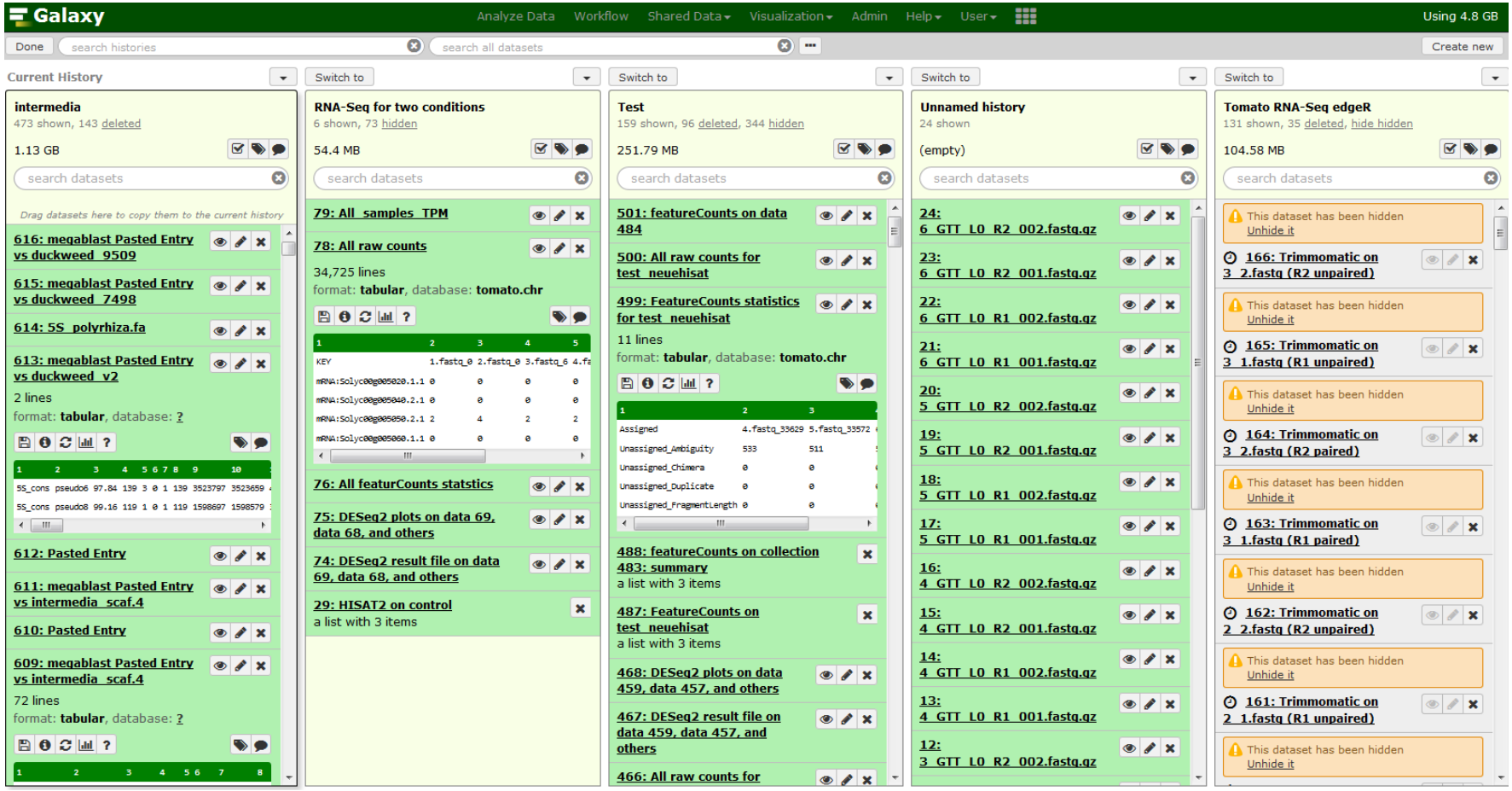
Multiple histories

- You can have as many histories as you want
  - Each history should correspond to a **different analysis**
  - Should have a meaningful **name**
- To see all of your histories, use the history switcher



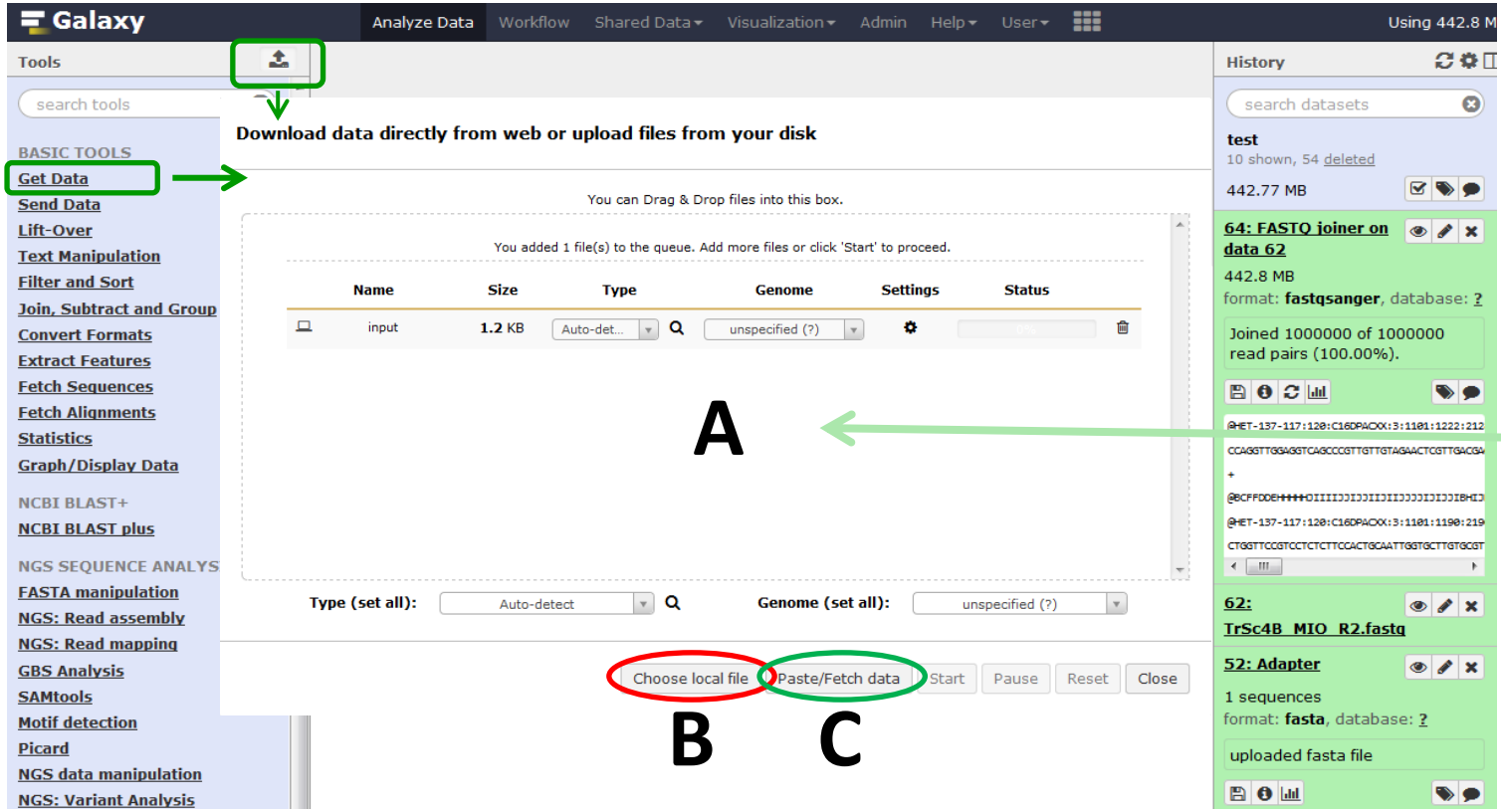
Multiple histories

- You can have as many histories as you want
  - each history should correspond to a **different analysis**
  - and should have a meaningful **name**



2 Data import

Data upload



Input data file

Go to **Get data** > **Upload file** **from your computer**:

A) Move your data file (<200mb) from your computers desktop into the white popup.

or

B) Choose the location of your files (preferred if your have multiple files to upload)

or

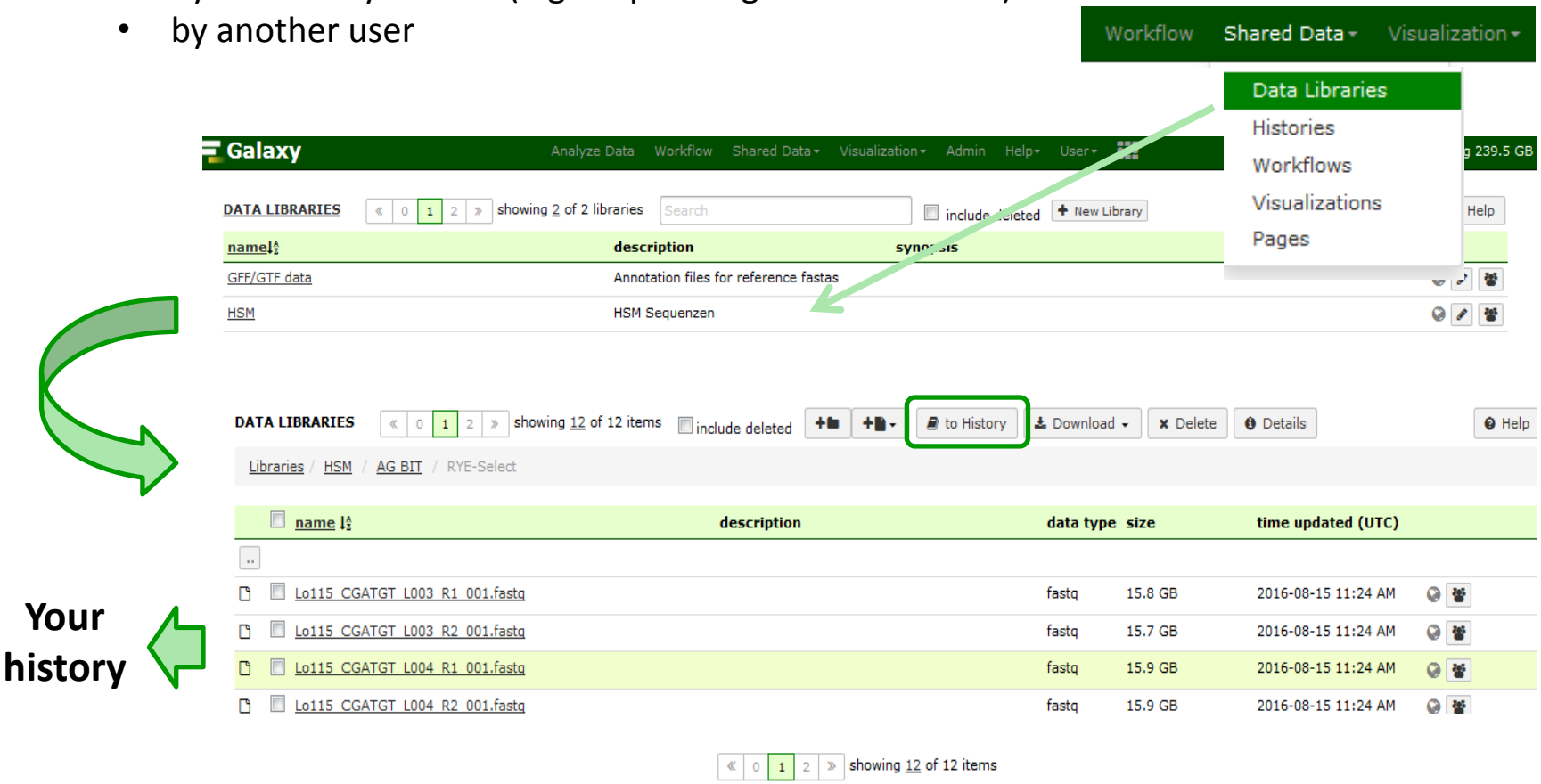
C) Paste data direcly using „Copy/Paste“ operations (keyboard, mouse)

Press the **Start** button to complete the import

Data libraries

Data libraries are a convenient framework within Galaxy to store and share data.
Data can be shared with you:

- by the Galaxy admins (e.g. sequencing data from HSM)
- by another user



Datatypes

- Tools only accept input datasets with the appropriate datatypes!
- When uploading a dataset, its datatype can be either:
  - automatically detected (fasta, txt, tabular)
  - assigned by user (fastq, other datatypes)
- Dataset produced by a tool: datatype assigned by the tool
- To change the datatype of a dataset in history:
  - *Edit Attributes* and *Datatype*
  - *Edit Attributes* and *Convert Formats* 

Datatypes

- Sequences (FASTA, FASTQ, ABI/SCF, SFF)
- Alignments (MAF, SAM/BAM, AXT, LAV)
- Intervals (BED, INTERVAL, GFF(3), WIG)
- Tabular data (e.g. CSV)
- Others (HTML, TXT, LPED)

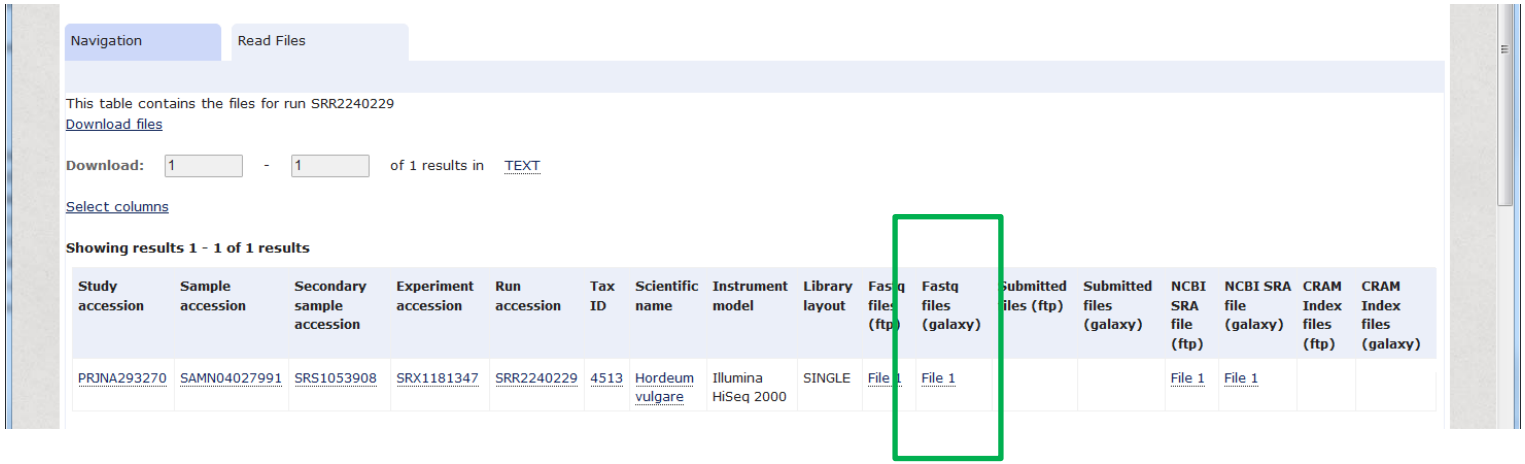
- Compressed file formats accepted (.gz, .zip)

- Format conversion
  - Converters included

- Not supported
  - MS Office binaries (Excel, Word) => export as „txt“ or „csv“

Data upload

- Private data
  - Upload from your own system (or copy/paste)
  - Import from shared data library (provided by the admins to keep restricted datsets confidential)
  - Private BLAST databases can included (also with ristriction to special users/groups at IPK)
- Public data
  - [Reference genomes from web services (*UCSC Genome Browser*)]
  - From URL (e.g. import from ENA SRA or NCBI SRA)

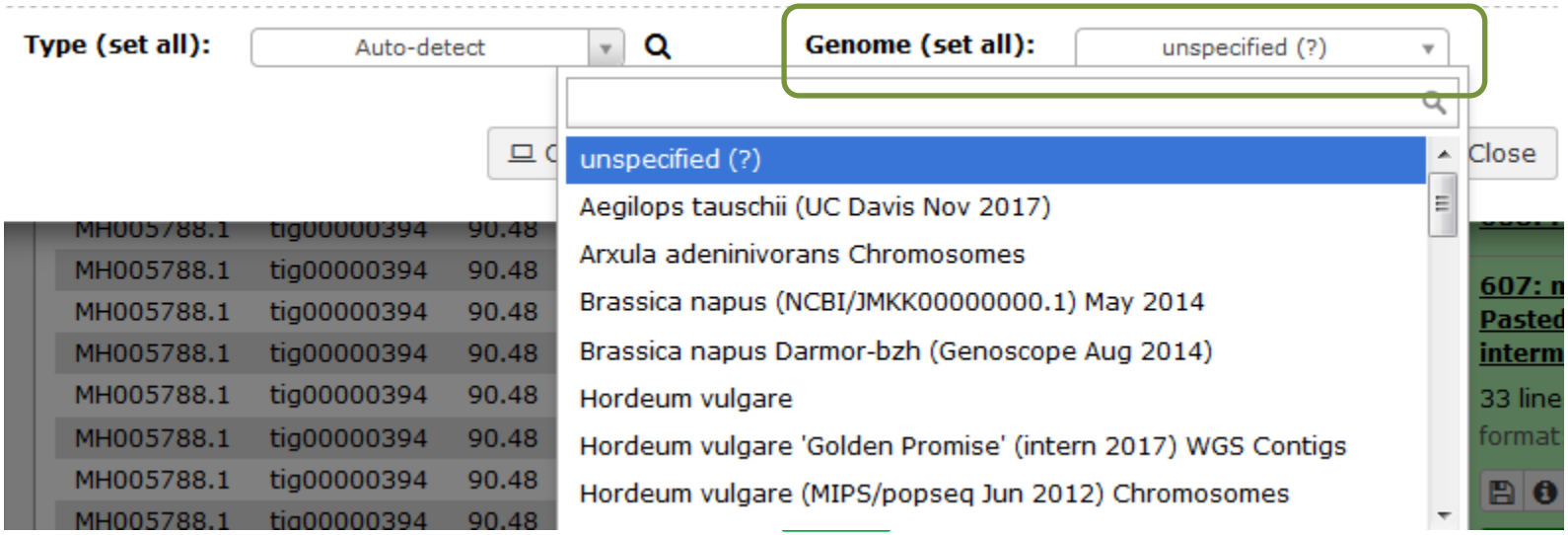


Local reference ressources

Reference genomes

- Genome build specifies which genome assembly a dataset is associated with
- Can be automatically detected or assigned by user
- New builds can be added by the admin

- Some tools allow to create indices from uploaded files (blast DBs)
  - Con: very time consuming for whole genome files (plants)



3 Running Galaxy

Processing your job

1

After submitting the job, a new box will be created, showing the job progress:

Job submitted

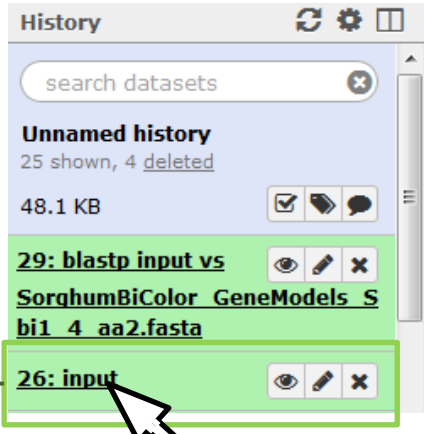
Job running

Job completed or Job failed

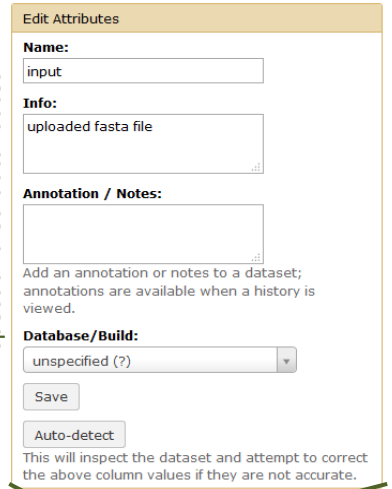
The color code indicates the job progress

2

Your history panel with the uploaded dataset „input“ (B)



Click to open



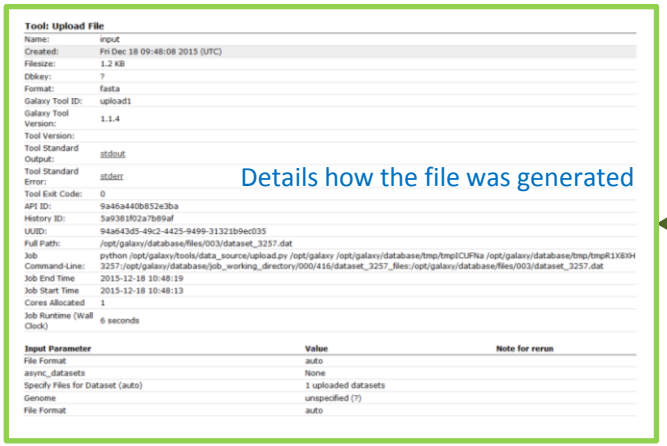
Inspect and edit attributes

3 View your data

Delete your data

Preview of the data

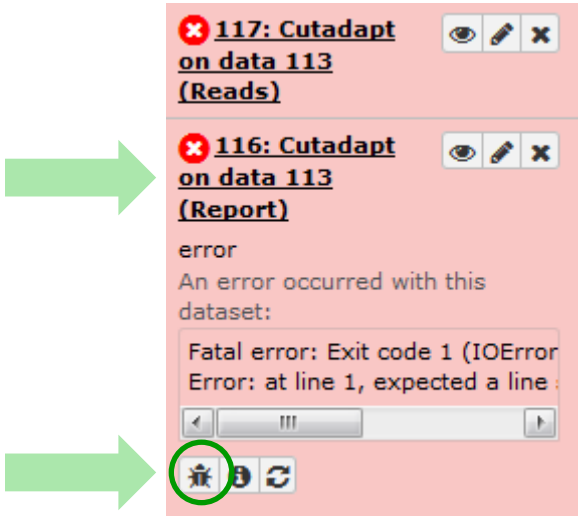
Download the file to your workspace



Re-run the job using the same configuration (or change some parameters)

Dealing with failed runs

- Check your input data
- Read the error report!



- Search for the error message
- https://biostar.usegalaxy.org
- Report the error to the local Galaxy administrators

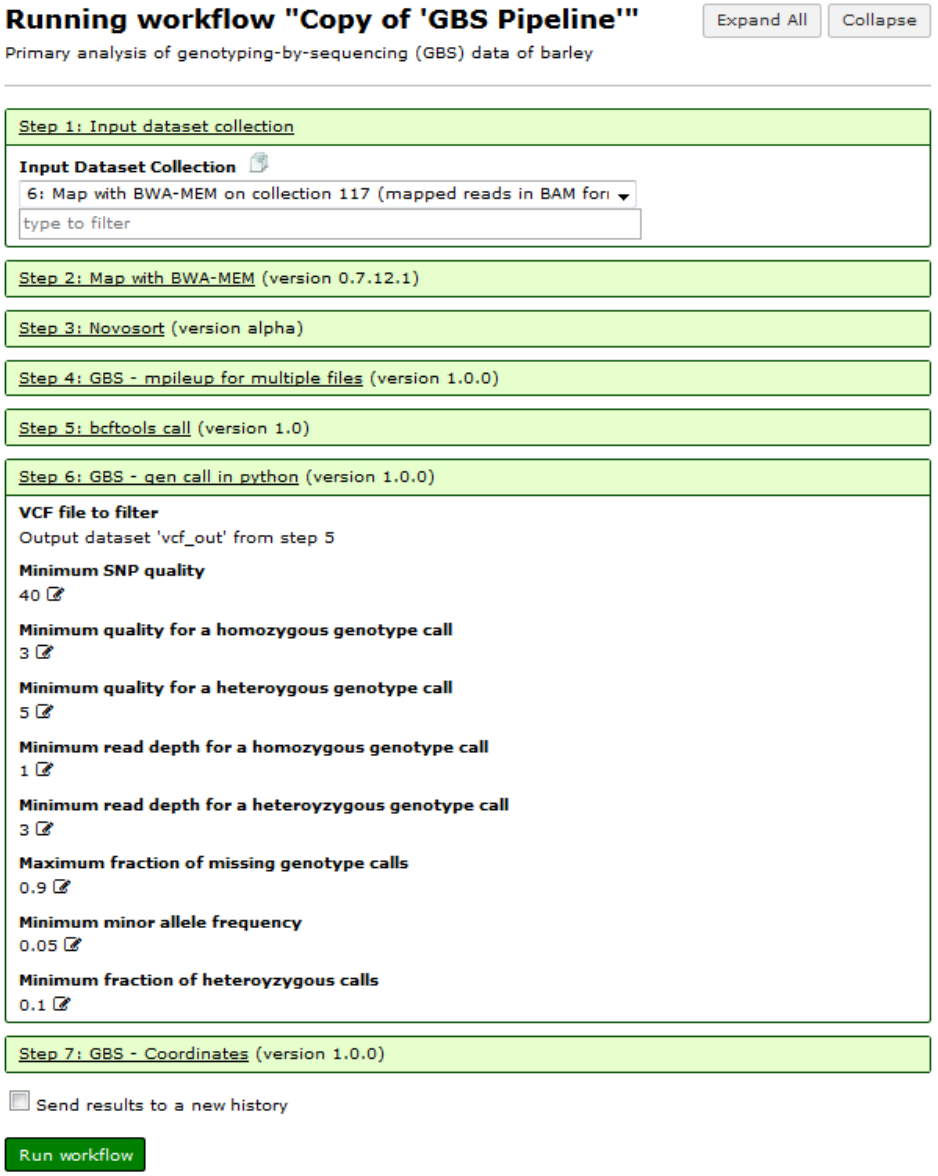
Biostars
GALAXY EXPLAINED

Workflows

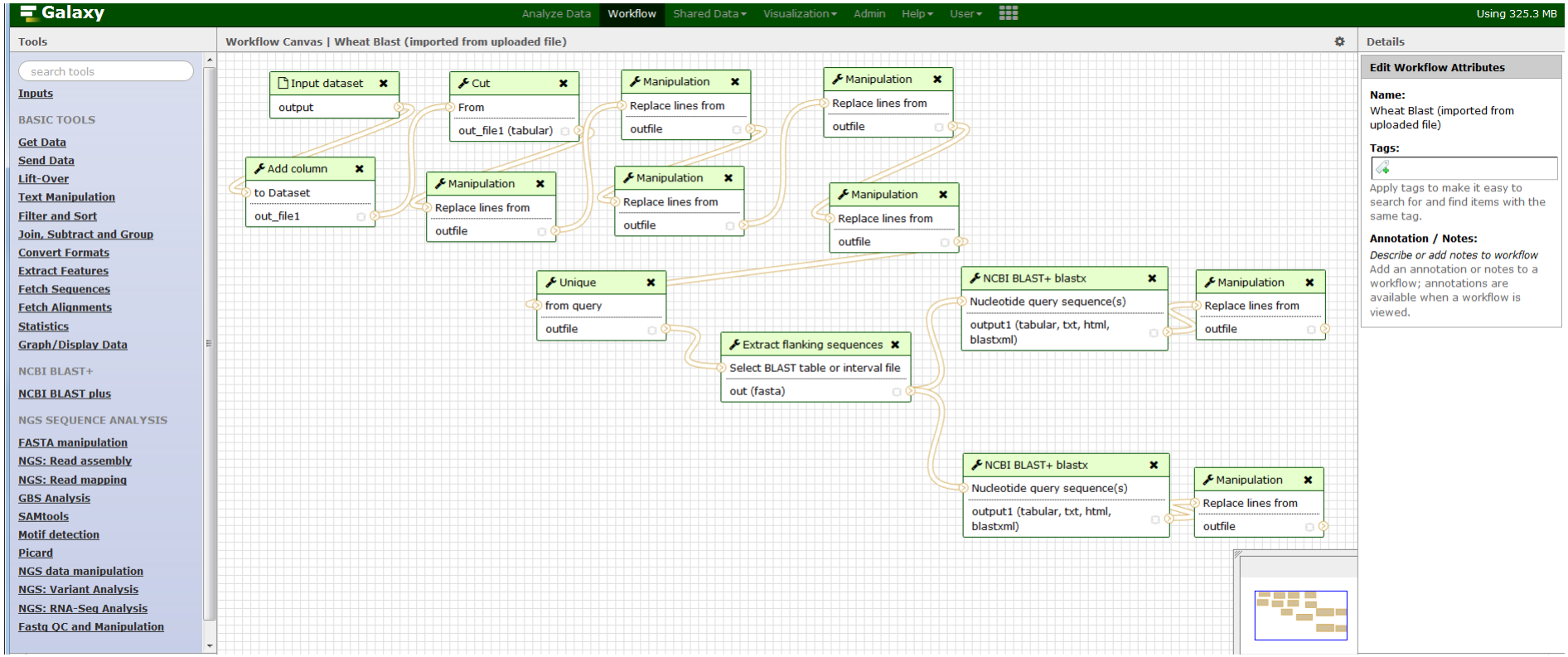
A workflow is ..

.. a series of tools and dataset actions that run in sequence as a batch operation.

- Workflows specify the steps in a process.
- Workflows are analysis that are meant to be run, each time with different user provided datasets.
- **Reproducible and well documented bioinformatic pipelines**



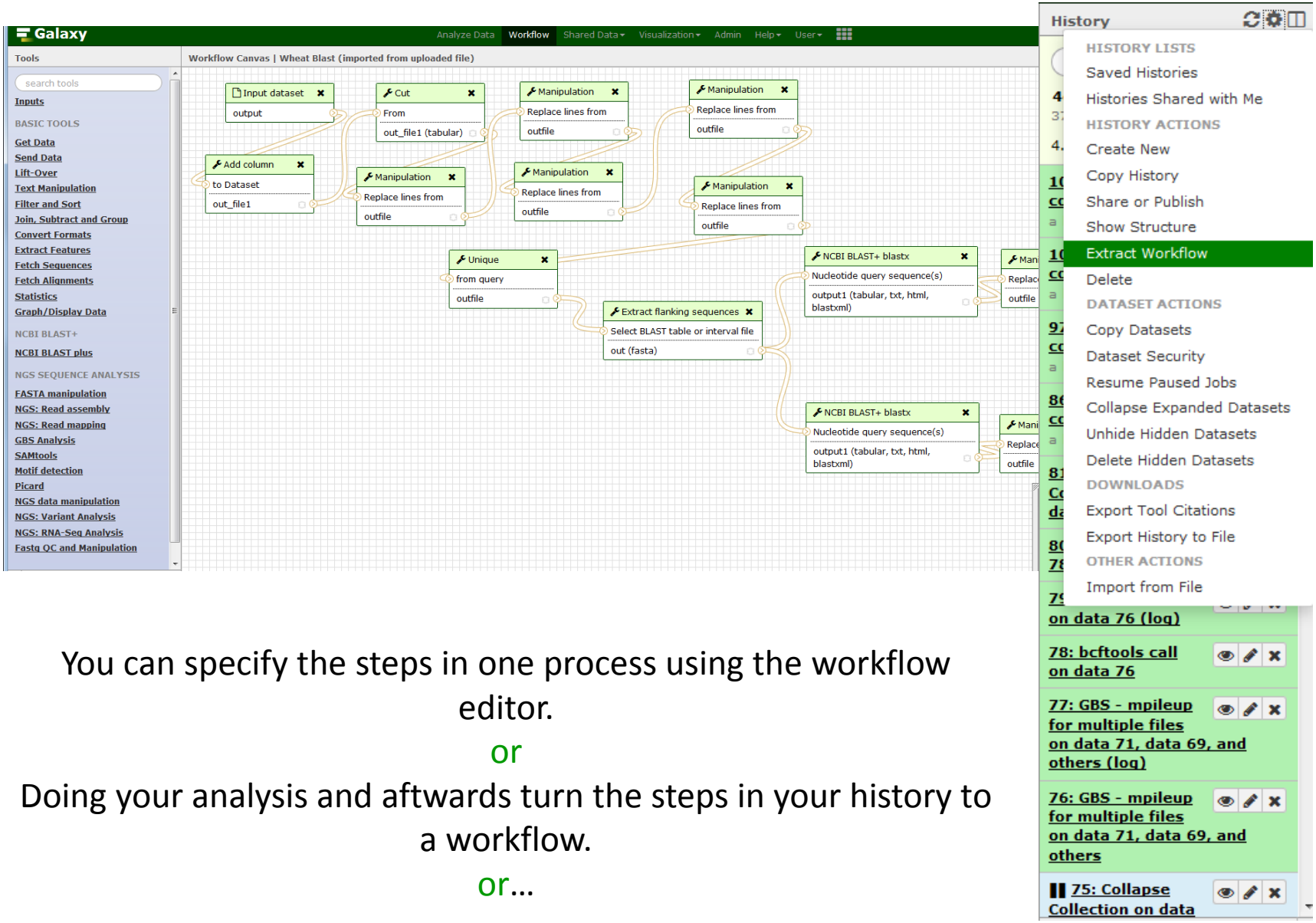
Preparing own workflows



Workflows can be created from scratch using the *workflow editor*

Or…

Preparing own workflows



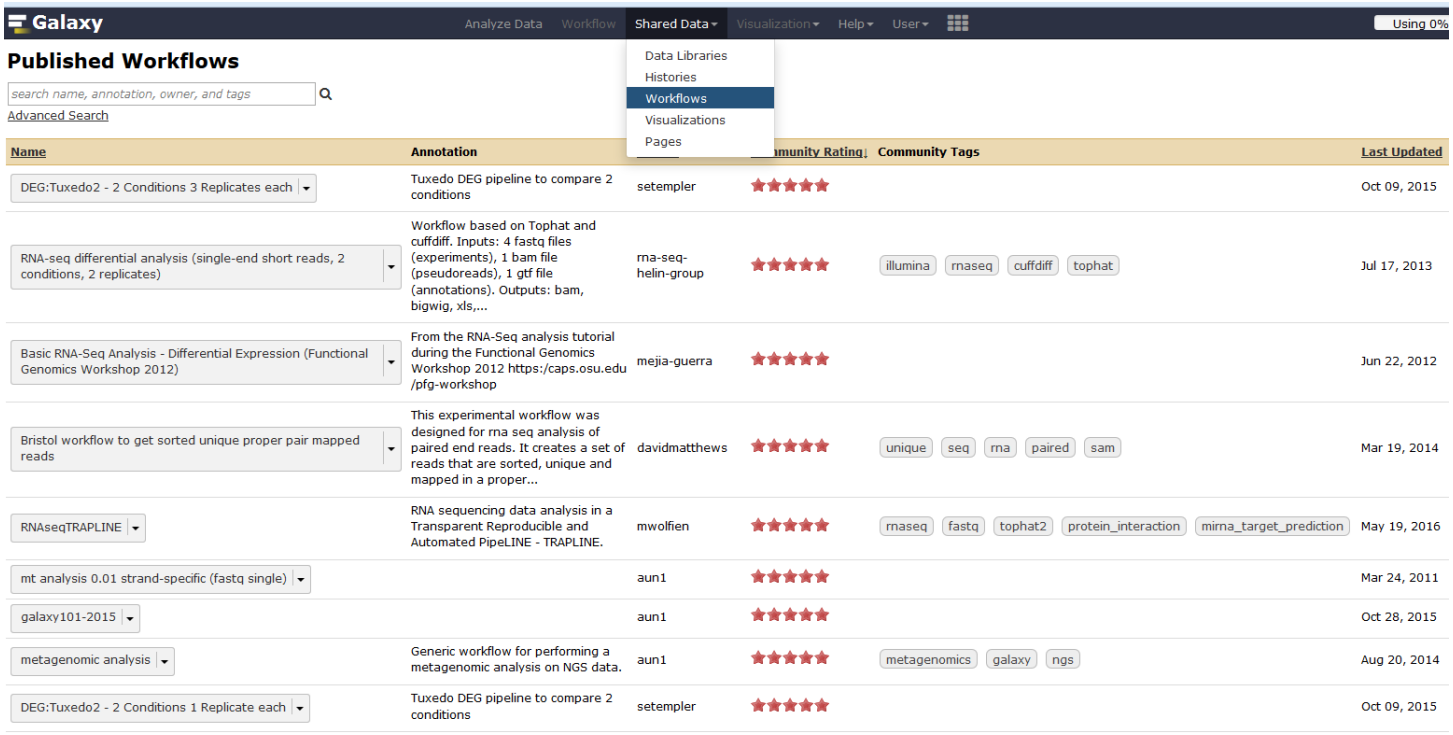
You can specify the steps in one process using the workflow editor.

or

Doing your analysis and aftwards turn the steps in your history to a workflow.

or…

Import workflows



Obtain public workflows from https://usegalaxy.org/ developed by the community.

or

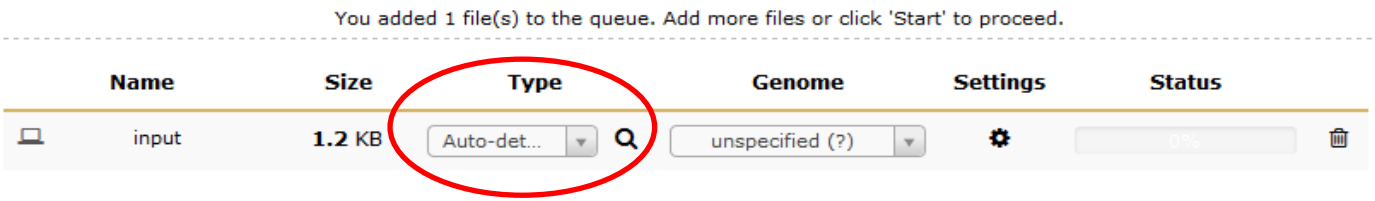
Share own workflows within the group or IPK.

Handling fastq files

By default Galaxy assign ‘fastq’ data type to uploaded FASTQ files.
In this case the offset is not specified, and many tools do not recognize the data.

- *fastqillumin*a: old illumina quality score encoding (offset 64, Illumina1.3+)
- *fastqsanger*: new Illumina1.8+ / Sanger quality score encoding

⇒Adjust format during upload



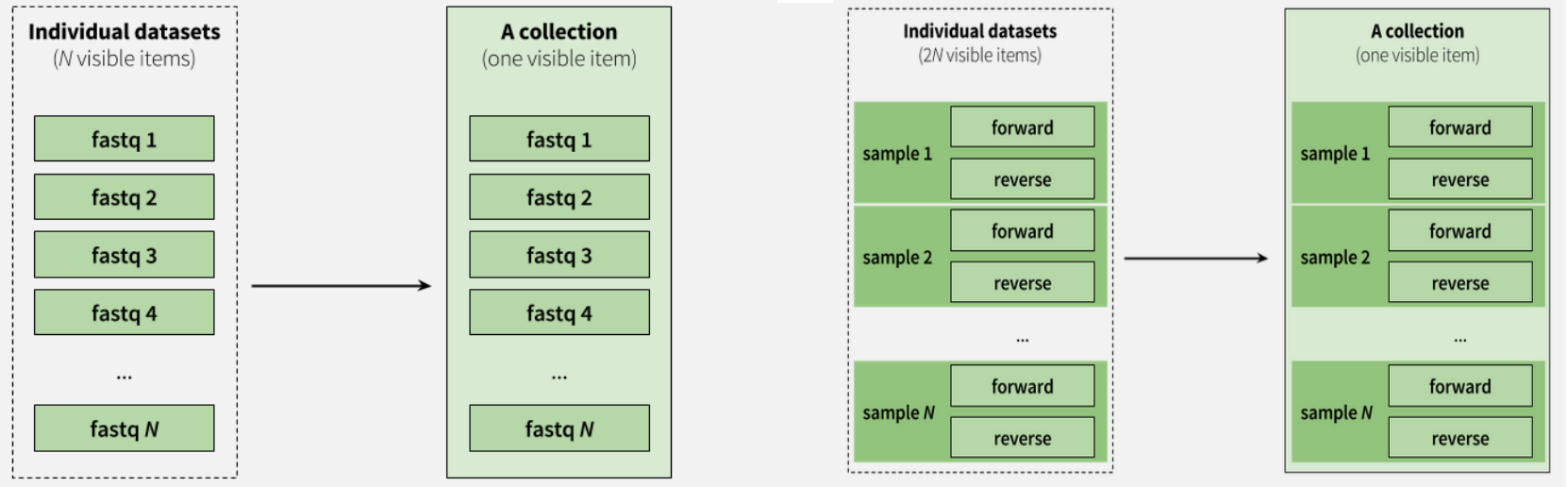
⇒Change data type in your history

Dataset collections

Collections allow users to handle and process a large number of samples at once.

Lists can contain a arbitrary number of elements

Paired list, containing a forward and reverse element each

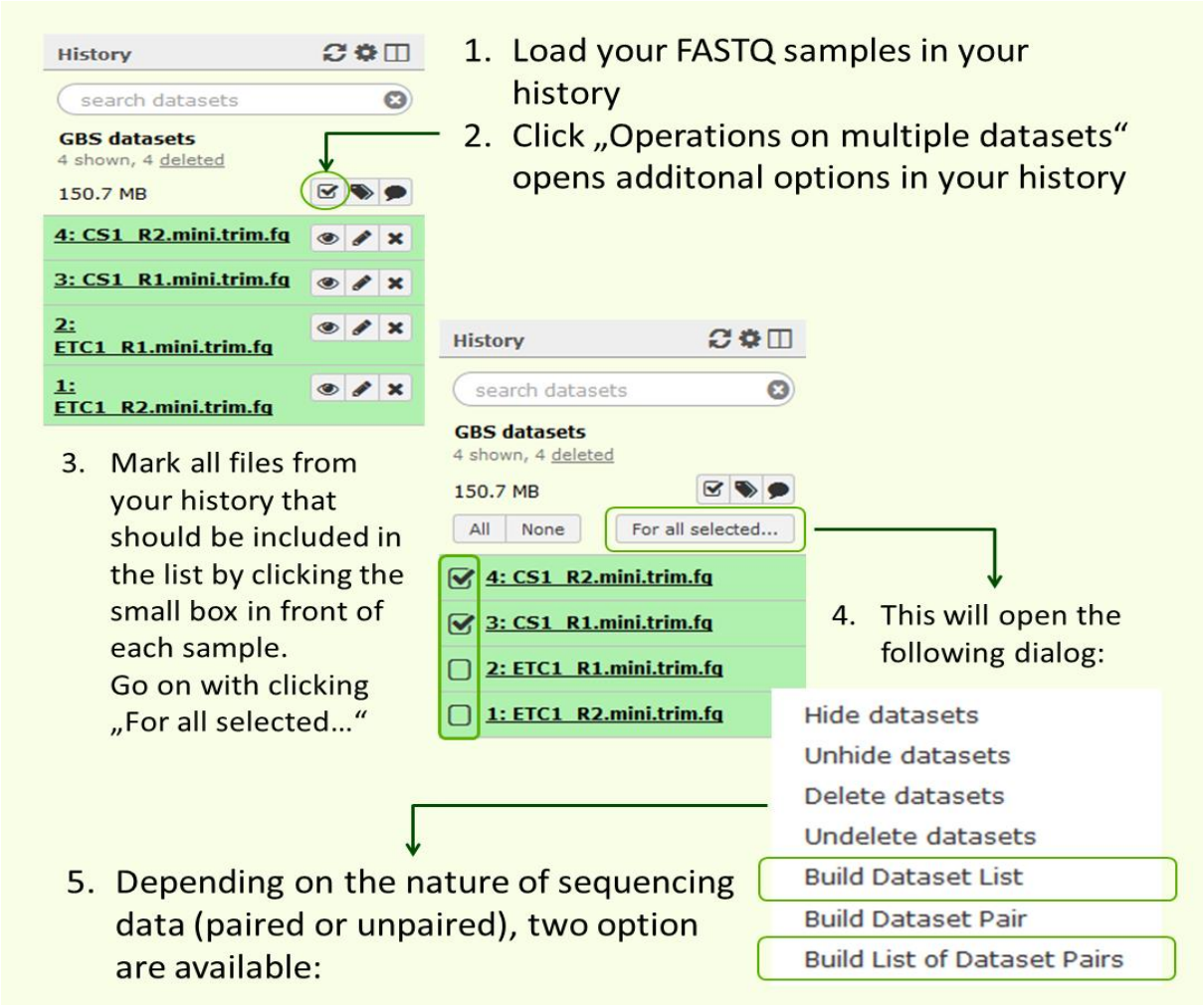


https://galaxyproject.org

Uploading data into collections directly to bypass the need to upload single datasets into history first.

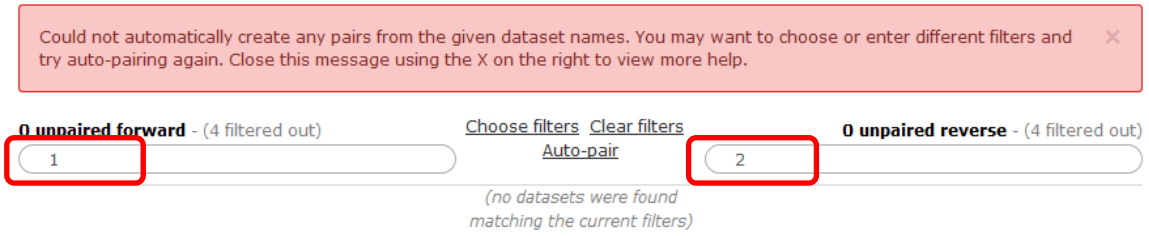
Dataset collections

Collections allow users to handle and process a large number of samples at once.

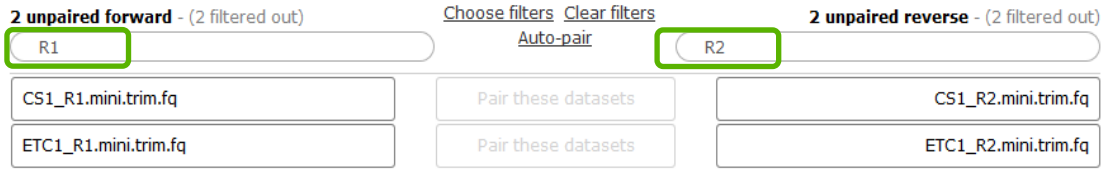


Paired dataset collections

Note: If your pairs are not named by „_1“ and „_2“ ending, you will probably see this message:

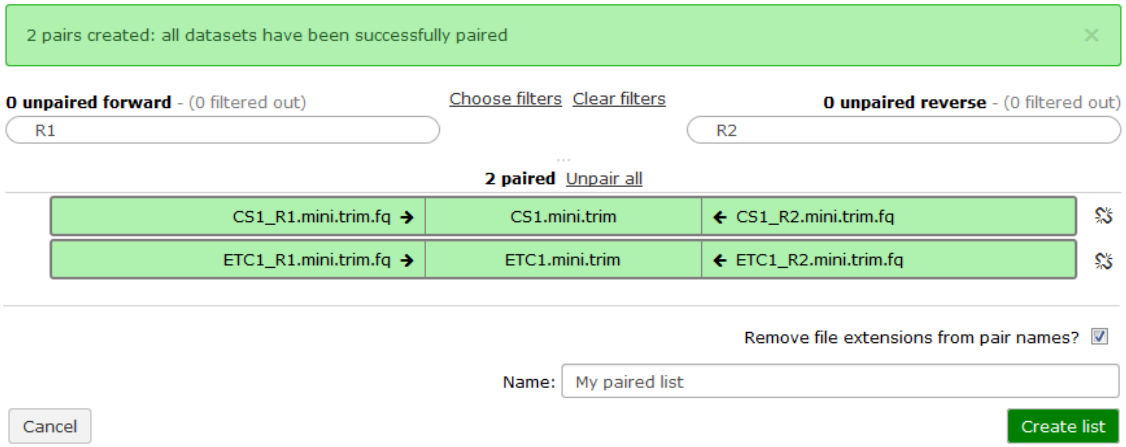


This should be corrected by typing the naming pattern of your data.
In this example „_R1“ and „_R2“:

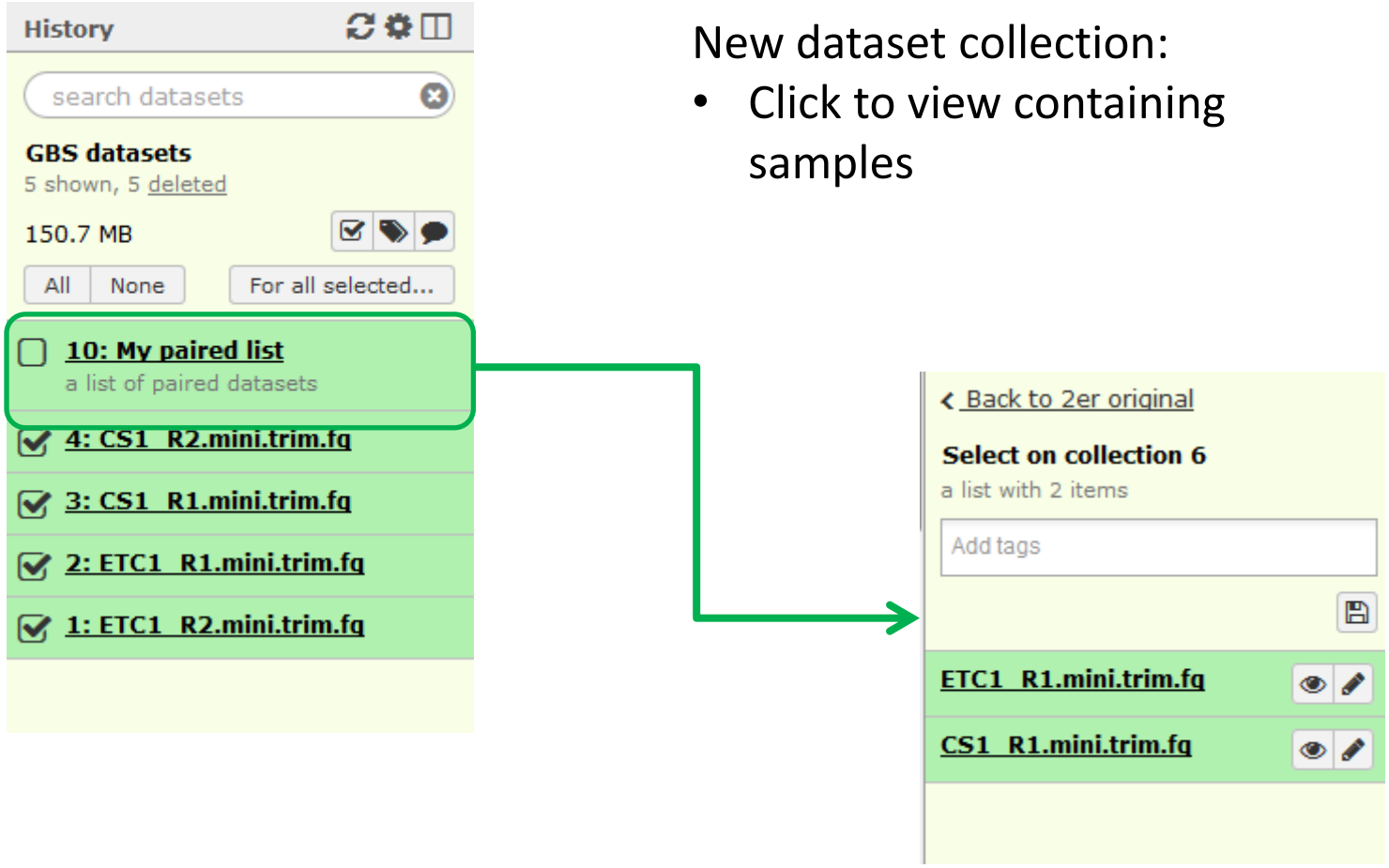


Check the correct pairing of the files. Then click „Auto-Pair“.

The files are paired according to their name scheme. Name and create the list now.



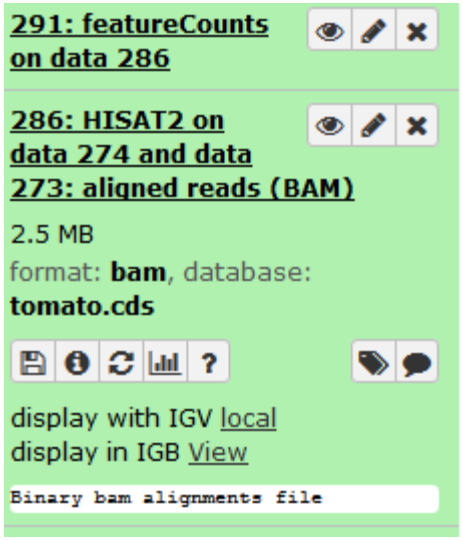
Paired dataset collections



New dataset collection:

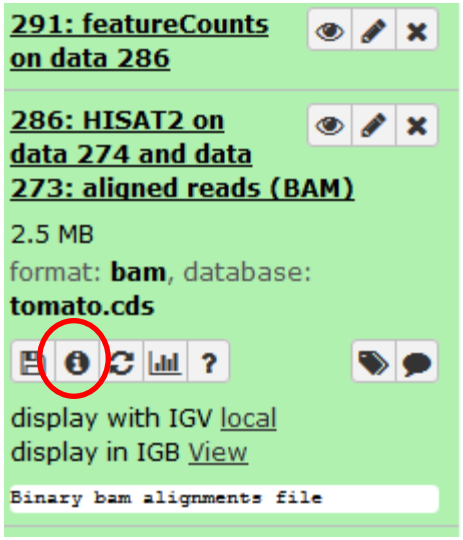
- Click to view containing samples

Logfiles output



Example: HISAT mapping - where can I find mapping statistics?

Logfiles output

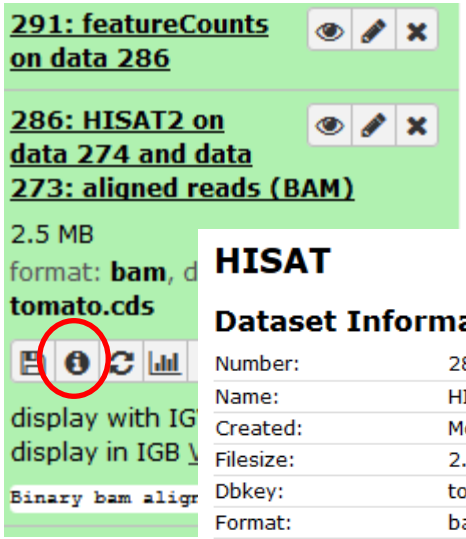


Example: HISAT mapping - where can I find the mapping statistics?

=> Use samtools flagstat

=> Check for additional output!

Logfiles output



Example: HISAT mapping - where can I find mapping statistics?

HISAT

Dataset Information

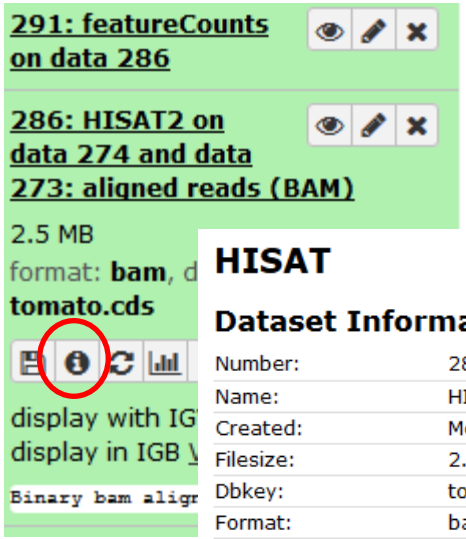
Number:	286
Name:	HISAT2 on data 274 and data 273: aligned reads (BAM)
Created:	Mon 18 Feb 2019 01:36:49 PM (UTC)
Filesize:	2.5 MB
Dbkey:	tomato.cds
Format:	bam

Job Information

Galaxy Tool ID:	toolshed.g2.bx.psu.edu/repos/iuc/hisat2/hisat2/2.1.0
Galaxy Tool Version:	2.1.0
Tool Version:	/opt/galaxy/dependency_dir/_conda/envs/mulled-v1-3b104c294f65450b09ba89d24826c61eddd553d789c9f1ae48a29715de1b7 version 2.1.0 64-bit Built on login-node03 Wed Jun 7 15:53:42 EDT 2017 Compiler: gcc version 4.8.2 (GCC) Options: -O3 -m64 -r -DPOPCNT_CAPABILITY Sizeof {int, long, long long, void*, size_t, off_t}: {4, 8, 8, 8, 8, 8}
Tool Standard Output:	stdout
Tool Standard Error:	stderr
Tool Exit Code:	0

Empty, the content is redirected to your BAM file

Logfiles output



Example: HISAT mapping - where can I find the mapping statistics?

HISAT

Dataset Information

Number:	286
Name:	HISAT2 on data 274 and data 273: aligned reads (BAM)
Created:	Mon 18 Feb 2019 01:36:49 PM (UTC)
Filesize:	2.5 MB
Dbkey:	tomato.cds
Format:	bam

Job Information

Galaxy Tool ID:	toolshed.g2.bx.psu.edu/repos/iuc/hisat2/hisat2/2.1.0
Galaxy Tool Version:	2.1.0
Tool Version:	/opt/galaxy/dependency_dir/_co version 2.1.0 64-bit Built on login -DPOPCNT_CAPABILITY Sizeof {in
Tool Standard Output:	stdout
Tool Standard Error:	stderr
Tool Exit Code:	0

```
19941 reads; of these:
  19941 (100.00%) were paired; of these:
    977 (4.90%) aligned concordantly 0 times
    18741 (93.98%) aligned concordantly exactly 1 time
    223 (1.12%) aligned concordantly >1 times
    ----
    977 pairs aligned concordantly 0 times; of these:
      118 (12.08%) aligned discordantly 1 time
    ----
    859 pairs aligned 0 times concordantly or discordantly; of these:
      1718 mates make up the pairs; of these:
        1253 (72.93%) aligned 0 times
        421 (24.51%) aligned exactly 1 time
        44 (2.56%) aligned >1 times
96.86% overall alignment rate
```

Acknowledgement

Special thanks to

- useGalaxy.eu team (especially Björn Grüning and Helena Rasche)
- BIT team
- IPK PostDoc board